

Utilizing Lexical Similarity for pivot translation involving resource-poor, related languages

Anoop Kunchukuttan, Maulik Shah, Pradyot Prakash, Pushpak Bhattacharyya

Center For Indian Language Technology

Department of Computer Science & Engineering

Indian Institute of Technology Bombay

{anoopk,maulik.shah,pradyot,pb}@cse.iitb.ac.in

Abstract

We investigate the use of pivot languages for phrase-based statistical machine translation (PB-SMT) between related languages with limited parallel corpora. We show that *subword-level pivot translation* via a related pivot language is: (i) highly competitive with the best direct translation model and (ii) better than a pivot model which uses an unrelated pivot language, but has at its disposal large parallel corpora to build the source-pivot (S-P) and pivot-target (P-T) translation models. In contrast, pivot models trained at word and morpheme level are far inferior to their direct counterparts. We also show that using *multiple related pivot languages* can outperform a direct translation model. Thus, *the use of subwords as translation units coupled with the use of multiple related pivot languages can compensate for the lack of a direct parallel corpus*. Subword units make pivot models competitive by (i) utilizing lexical similarity to improve the underlying S-P and P-T translation models, and (ii) reducing loss of translation candidates during pivoting.

1 Introduction

Related languages are those that exhibit lexical and structural similarities on account of sharing a **common ancestry** or being in **contact for a long period of time** (Bhattacharyya et al., 2016). Machine Translation between related languages is a major requirement since there is substantial government, commercial and cultural communication among people speaking related languages (Europe, India and South-East Asia being prominent examples and linguistic regions in Africa pos-

sibly in the future). These translation requirements generally fall into the following scenarios: [A] between related languages, [B] to/from a *lingua franca* like English. However, most of these language pairs have few or no parallel corpora.

When limited parallel corpus is available, translation using subword units like characters, orthographic syllables, *etc.* is useful in scenario [A] (Vilar et al., 2007; Kunchukuttan and Bhattacharyya, 2016c). This method utilizes *lexical similarity* between the related languages. Lexical similarity refers to sharing of many words with similar form (spelling/pronunciation) and meaning among related languages *e.g. blindness* is *andhapana* in Hindi, *aandhaLepaNaa* in Marathi. On the other hand, if no parallel corpus is available between two languages, pivot based SMT provides a systematic way of using a third language (called the *pivot language*) with which the source and target languages share parallel corpora. The general pivot based approach does not make any assumptions about relatedness between the source, pivot and target languages.

In this paper, we propose methods for translation involving related languages when no parallel corpus is available. For scenario [A] (translation between related languages), we utilize lexical similarity between the languages to improve pivot translation by: (i) using a pivot language that is related to source and target languages, (ii) using a subword translation unit *viz.* orthographic syllables (OS), (iii) identifying the pivot language based on its similarity with the source and/or target language. The contributions of our work are:

We show that using a **pivot language related to the source and target languages along with subword level translation** (with OS as translation unit) is very competitive with the best direct translation systems. Specifically, we show that our ap-

proach:

- significantly outperforms morpheme and word level pivot translations (about 15% and 60% increase in BLEU score respectively).
- is competitive with a skyline OS level direct translation model (achieving about 90% of direct system’s BLEU score) and better than the corresponding word-level direct translation model.
- is better than a word level pivot translation model which uses an unrelated pivot language and is trained on a large parallel corpus.
- is more robust to change of translation domain as compared to word and morpheme level pivot models.

We show that **use of multiple pivot languages** can compensate for the lack of direct parallel corpora. Specifically,

- We show that combining pivot systems using different pivot languages can outperform the best bilingual direct translation system. Thus multilinguality acts as booster to deliver the best translation results between related languages.
- Our investigations on the choice of pivot language suggests that a pivot language closer to the target language may be preferable for better translation performance.

To the best of our knowledge, ours is the first work that shows that **a pivot system can be very competitive with (or even outperform) a direct system (in the restricted case of related languages)**.

The rest of the paper is organized as follows. Section 2 discusses related work. Since we utilize lexical similarity between languages we discuss this concept in Section 3. Section 4 explains our method for translation between related languages. Section 5 describes our experimental setup.

The next two sections analyse the results of our experiments on pivot translation between related languages: Section 6 discusses the effect of subword units, while Section 7 discusses the effect of multiple pivot languages and factors affecting choice of pivot languages. Section 8 analyzes the results of the our initial investigations for translation between unrelated languages via a pivot which is related to either source/target (but not both) (scenario B). Finally, we summarize our work and discuss future work in Section 9.

2 Related Work

Our work straddles two strands of research in statistical machine translation: (i) subwords as basic units for translation between related languages, and (ii) pivot-based machine translation.

Modeling the lexical similarity among related languages is the key to building good-quality SMT systems with limited parallel corpora. This can be achieved by subword level transformations. Lexical similarity can be modelled in the standard SMT pipeline by transliteration of words while decoding (Durrani et al., 2010) or post-processing (Nakov and Tiedemann, 2012; Kunchukuttan et al., 2014). An alternative method to improve translation quality is to use subwords as basic translation units. Subword units like character (Vilar et al., 2007; Tiedemann, 2009), character n-gram (Tiedemann and Nakov, 2013), orthographic syllables (Kunchukuttan and Bhattacharyya, 2016c) and Byte Pair Encoded units (Kunchukuttan and Bhattacharyya, 2016b) have been explored and have been shown to improve translation quality to varying degrees. We have used orthographic syllables as the basic translation unit in our experiments, since they have been shown to outperform other subword units.

Pivot translation provides a systematic way for translation between two languages through an intermediate language. Multiple approaches to pivoting have been proposed viz. (i) synthetic corpus generation (Adri’a De Gispert, 2006) (ii) transfer-based/pipelining (Utiyama and Isahara, 2007) (iii) phrase-table triangulation (Utiyama and Isahara, 2007; Wu and Wang, 2007; Cohn and Lapata, 2007). We use triangulation in our experiments as it has been shown to outperform other approaches.

In the context of translation involving related languages, there has been some work using subword level units for pivot-based SMT. Character-based pivot translation has been explored for one leg of the pivot system (S-P or P-T) when the pivot is related to either the source or target (but not both) (Tiedemann, 2012; Tiedemann and Nakov, 2013). In our scenario, the source, target and pivot languages are all related. Hence subword level translation is possible for both legs (source-pivot and pivot-target). Morpheme level pivot translation has been shown to be better than word level pivot translation (More et al., 2015). More et al.

(2015) and Dabre et al. (2015) have experimented with multiple pivot languages, but the translation accuracy is lower than the direct model.

3 Lexical Similarity in related languages

Two words are said to be lexically similar if they have similar form (spelling/pronunciation) and meaning *e.g.* *time* is *samay* in Hindi, *samayam* in Malayalam. These words could be cognates, lateral borrowings or loan words from other languages. Two languages are said to be lexically similar if they share a lot of lexically similar words. Lexical similarity is a key characteristic of related languages.

*Ethnologue*¹ computes the percentage of lexical similarity between two linguistic varieties by comparing a set of standardized wordlists and counting those forms that show similarity in both form and meaning (Rensch, 1992). However, this requires extensive fieldwork and linguistic analysis.

For our analysis, we use a simpler measure of lexical similarity which depends on the orthography, rather than phonology. For languages which use scripts with a high grapheme-to-phoneme correspondence, this is a reasonable approach to take. We use the Longest Common Subsequence Ratio (LCSR) as a measure of lexical similarity between two strings (Melamed, 1995):

$$LCSR(s_1, s_2) = \frac{|LCS(s_1, s_2)|}{\max(|s_1|, |s_2|)} \quad (1)$$

where, s_1, s_2 are two strings and *LCS* is the longest common subsequence between them.

LCSR can be a versatile tool to analyze related languages and can be applied to different linguistic entities to measure their lexical similarity. In addition to its utility for measuring lexical similarity between **words** (Melamed, 1999; Kondrak, 2005), it can also be used to compare **sentences** (assuming the two sentences to be a sequence of characters). The sentences could be from the *same language* (*e.g.* reference translation and output of MT system (Tiedemann, 2012)) or *different languages* (*e.g.* parallel translations (Kunchukuttan and Bhattacharyya, 2016c)). To compute LCSR across languages, the script should be either the same or a correspondence between characters should exist. The lexical similarity between two **languages** can also be computed by averaging over the lexical similarities of sentence

	Indo-Aryan						Dravidian	
	pan	hin	ben	guj	mar	kok	tel	mal
pan								
hin	68.0							
ben	44.6	52.3						
guj	52.0	58.9	49.0					
mar	47.2	53.0	46.4	54.1				
kok	42.2	46.7	41.6	48.5	54.5			
tel	39.1	42.9	38.8	41.8	41.1	39.2		
mal	30.4	33.2	32.1	32.9	33.6	32.1	39.2	
tam	27.1	29.0	28.0	28.0	29.6	29.0	35.5	39.0

Table 1: Lexical Similarity between Indian languages used in experiments. ISO 639-3 language codes shown.

pairs in a parallel corpus. Table 1 shows the lexical similarities between various Indian languages which we have used in our experiments. These were computed on the multilingually aligned ILCI parallel corpus (Jha, 2012). The pairwise rankings of languages as measured by LCSR agree with the general consensus about similarity between Indian languages.

4 Pivot Translation for Related Languages

We first train phrase-based SMT models between S-P and P-T language pairs with subword units (orthographic syllables in our case). We create a pivot translation system by combining the S-P and P-T models using phrase table triangulation. If multiple pivot languages are available, linear interpolation is used to combine pivot translation models. In this section, we describe each component of our system and the design choices.

4.1 Orthographic Syllable level translation

We use *orthographic syllables* (OS) (Kunchukuttan and Bhattacharyya, 2016c) as the basic units of translation. It is a linguistically-motivated, variable-length unit which consists of a consonant core with zero or more vowels (a C+V combination) (*e.g.* *spacious* would be segmented as *spa ciou s*). Since the vocabulary is much smaller than the morpheme and word level models, data sparsity is not a problem. The variable length units provide appropriate context for translation between related languages.

This unit has outperformed character n-gram, word and morpheme level models for the task of translation between related languages. Or-

¹www.ethnologue.com

thographic syllables outperform other units even when: (i) the languages are not very *closely* related (ii) the languages do not have a genetic relation, but only a contact relation.

Using OS level models for pivot translation ensures that the underlying S-P and P-T translation models are better than the corresponding models trained on other translation units.

4.2 Using a related pivot language via Triangulation

We use phrase-table triangulation to fuse the src-pivot and pivot-tgt OS level models for generating the pivot model’s phrase table. Triangulation joins the two tables on the common phrases in the pivot language and recomputes the probabilities in the phrase table (direct/inverse phrase and lexical translation probabilities) by assuming a generative process and making a few independence assumptions:

$$P(\bar{s}|\bar{t}) = \sum_{\bar{p}} P(\bar{s}|\bar{p}, \bar{t}) P(\bar{p}|\bar{t}) \quad (2)$$

$$\approx \sum_{\bar{p}} P(\bar{s}|\bar{p}) P(\bar{p}|\bar{t}) \quad (3)$$

where \bar{s} , \bar{p} and \bar{t} are source, pivot and target language phrases respectively

As compared to word and morpheme level models, OS level pivot models are likely to find more common pivot language phrases because of the smaller vocabulary size. Hence, data sparsity, a problem recognized by Dabre et al. (2015) and More et al. (2015), will be a lesser impediment to pivoting for OS level models.

4.3 Multiple Pivot Languages

When multiple pivot languages are available, we can either use all pivot languages or choose the best pivot language.

The choice of pivot language is affected by its relatedness to the source and target language (Paul et al., 2013). We studied how lexical similarity of the pivot language to the source and/or target language affects translation quality.

We also experimented with combining multiple pivot language translation models using linear interpolation (Bisazza et al., 2011). Linear interpolation assigns weights to each phrase table and the feature values for each phrase pair are interpolated

using these weights:

$$P(\bar{s}|\bar{t}) = \sum_i \alpha_i P_i(\bar{s}|\bar{t}) \quad (4)$$

subject to: $\sum_i \alpha_i = 1, \alpha_i \geq 0$

where, α_i = interpolation weight for phrase table i .

We experimented with different strategies for determining the interpolation weights using linguistic similarity between the languages.

5 Experimental Setup

This section describes languages and datasets used in our experiments and details of our system.

5.1 Languages

We experimented with multiple languages from the two major language families of the Indian sub-continent (Indo-Aryan branch of Indo-European and Dravidian). The Indian subcontinent is considered a *linguistic area* (Emeneau, 1956) due to convergence of linguistic properties as a result of contact between languages over a long period of time. Specifically, there is substantial overlap between the vocabulary of these languages to varying degrees due to cognates, language contact and loan-words from Sanskrit and English. All these languages have a rich inflectional morphology with Dravidian languages (and Marathi to some degree), being agglutinative. Table 1 shows the languages involved and their lexical similarities.

5.2 Dataset

We used the multilingual Indian Language Corpora Initiative (ILCI) corpus² for our experiments (Jha, 2012), containing sentences from tourism and health domains. The corpus contains sentences aligned across 11 languages. This multilingual alignment enables a fair comparison of (i) direct and pivot translation systems, (ii) multiple pivot languages, and (iii) lexical similarity of languages.

The data split is as follows – *training*: 44,777, *tuning* 1K, *test*: 2K sentences. Language models for word-level systems were trained on the target side of training corpora plus monolingual corpora from various sources [hin: 10M (Bojar et al., 2014), tam: 1M (Ramasamy et al., 2012), mar: 1.8M (news websites), mal: 200K, ben: 400K, pan: 100K (Quasthoff et al., 2006) sentences]. We

²available on request from TDIL (tdil-dc.in)

Lang Triple	Word	Morph	OS
mar-guj-hin	30.23	36.49	39.81
mar-hin-ben	16.63	21.04	22.92
mal-tel-tam	4.55	6.19	7.19
tel-mal-tam	5.13	8.29	8.39
hin-tel-mal	5.29	8.32	9.67
mal-tel-hin	10.03	13.06	17.26

Table 2: Compare different subwords (%BLEU). Lang triple refers to the source-pivot-target language. Scores in **bold** indicate highest values for the language triple.

used the target language side of the parallel corpora for morpheme and OS level LMs.

5.3 System details

PBSMT systems were trained using the *Moses* system (Koehn et al., 2007), with *mgiza*³ for alignment, the *grow-diag-final-and* heuristic for symmetrization of word alignments, and Batch MIRA (Cherry and Foster, 2012) for tuning. Cube pruning with *pop-limit=1000* was used for decoding (Kunchukuttan and Bhattacharyya, 2016a). We trained 5-gram LMs with Kneser-Ney smoothing for word and morpheme level models and 10-gram LMs for orthographic syllable level models.

We use unsupervised morphological segmenters trained with *Morfessor* (Virpioja et al., 2013) for obtaining morphemes. These segmenters were trained on the ILCI corpus and the Leipzig corpus (Quasthoff et al., 2006). We use the *tmtriangulate* for phrase-table triangulation⁴. Prior to triangulation, we prune phrase pairs with direct phrase translation probability less than 0.01 to make the triangulation process more efficient. We use the *combine-ptables* (Bisazza et al., 2011), (part of the *Moses* distribution), for linear interpolation of phrase tables.

6 Discussion: Subword units for pivot translation

In this section, we present results related to different subword units and analyse them.

Lang Triple	Pivot	Direct		
	OS	Word	Morph	OS
mar-guj-hin	39.81	38.87	42.81	43.69
mar-hin-ben	22.92	21.31	23.96	23.53
mal-tel-tam	7.19	6.52	7.61	7.84
tel-mal-tam	8.39	9.58	10.61	10.52
hin-tel-mal	9.67	8.49	9.23	10.46
mal-tel-hin	17.26	15.23	17.08	18.44

Table 3: Comparison of pivot and direct translation (%BLEU)

6.1 Comparison of different subword units

Table 2 compares pivot-based SMT systems built with different units. We experimented with language triples over various combinations of the language families (Indo-Aryan and Dravidian).

We observe that in each case **the OS level pivot model significantly outperforms word and morph-level pivot models** (an average improvement of about 61% and 14% improvement respectively). The OS level models show the greatest improvement over other units when the source and target languages belong to different families, showing that OS level models can utilize the lexical similarity between these languages. Translation between agglutinative Dravidian languages also shows a major improvement.

OS level pivot models are better than other units for two reasons. One, the underlying S-P and P-T translation models are better (average 16% and 3% improvement over word and morph-level models). However, this alone does not explain the substantial improvement in OS level pivot translation. The triangulation process, which is a join on common keys, is faced with sparsity arising from the large word and morpheme phrase-table vocabulary. The word level triangulated table actually loses translation candidates because of sparsity. On the other hand, **the OS phrase table vocabulary is smaller, so the impact of sparsity is limited** and the triangulated phrase table size increases by a few multiples for OS level models.

System	Google	OS-level
<i>Pivot</i>	<i>eng</i>	<i>tel</i>
hin-mal	4.19	5.96
mal-hin	7.92	11.33
mal-tam	2.28	5.82

Table 4: Related vs. unrelated pivot (%BLEU)

6.2 Comparison of pivot models with direct models

We also compared the OS-level pivot system with direct system trained on different translation units (See Table 3). It outperforms a word-level direct translation system between source and target by 5%, which is encouraging. Even more remarkable is that the OS level pivot model is competitive with the morph and OS level direct translation models (achieving about 95% and 91% of their respective BLEU scores). To put this fact in perspective, the BLEU scores of morph and word-level pivot systems are far below their corresponding direct systems (about 15% and 35% respectively).

These observations strongly suggests that pivoting at the OS level can reconstruct the direct translation system better compared to the word and morph level pivot systems, and the resultant OS level pivot systems are quite close to the best direct translation systems.

6.3 Using an unrelated pivot language

In the experiments described so far, we have considered a related pivot language with which the source and target language share small parallel corpora. We compare this to a very likely pivot scenario - the pivot language is an unrelated language like English with which the source and target languages share a lot of parallel corpora.

For this experiment, we used *Google Translate*⁵ as a translation system using an unrelated pivot. It is known that *Google Translate* uses English as a pivot language for many translation pairs where a direct translation corpus is not available (TAUS, 2013). For the languages we experimented with, this can be attested by the fact that English words turn up in the translations provided by *Google Translate*. *Google Translate* is presumably trained on large corpora, certainly orders of

Lang Triple	Pivot		Direct	
	Morph	OS	Morph	OS
hin-tel-mal	4.72	5.96	5.99	6.26
mal-tel-hin	8.29	11.33	11.12	13.32
mal-tel-tam	4.41	5.82	5.84	5.88

Table 5: Cross domain translation for different subwords (%BLEU)

magnitude larger than our translation systems, but at the word/morpheme level.

We compared our pivot translation models (using a related pivot language – Telugu –, trained on tourism and health domains) with Google Translate (using an unrelated pivot – English –) by testing on an agriculture domain test set of 1000 sentences from the ILCI corpus. This ensures a fair comparison between the two systems. Table 4 shows the Google Translate results alongside our pivot translation results. **The word/morpheme level translation using an unrelated pivot and large corpora is inferior to an OS level system using a related pivot and trained on small parallel corpora.**

For the language pairs involving *mal* as source language, we see that the Google Translate system produces a large number of OOVs. For the *hin-mal* language pair, the output of Google Translate was slightly shorter than reference translation, and the precision and recall were lower as compared to the OS-level model (as measured using METEOR).

6.4 Cross-Domain Translation

We also investigated if the OS level pivot models are robust to domain change by evaluating the translation models trained on tourism & health domains on an agriculture domain test set of 1000 sentences (from the ILCI corpus). Table 5 shows the results of these experiments. In this cross-domain translation scenario too, the OS level pivot models outperforms morph level pivot models, and is equivalent to a direct morph level model. The OS-level models systems experience much lesser drop in BLEU scores *vis-a-vis* direct models, in contrast to the morph level models. Since morph-level pivot models encounter unknown vocabulary in a new domain, they are less resistant to domain change than OS level models.

³github.com/amos-sm/mgiza

⁴github.com/tamhd/MultiMT

⁵<https://translate.google.com>

Weighting	mar-ben	mal-hin
equal	23.69	19.12
source	23.59	19.11
target	23.67	18.96
average	23.81	18.98
best pivot	22.92 (<i>hin</i>)	17.52 (<i>tel</i>)
direct	23.53	18.44

Table 6: Combination of Multiple Pivots (%BLEU). Pivots for (i) mar-ben: guj, hin, pan (ii) mal-hin: tel, mar, guj.

7 Discussion: Multiple pivot languages

We discuss the results of experiments with multiple pivots and study the choice of pivot language. The experiments in this section refer to OS level pivot models unless otherwise specified.

7.1 Combining Multiple Pivot Models

We investigated if multiple pivot translation systems can act as a substitute for a direct translation system. For this we combined multiple pivot translation systems using linear interpolation. We tried various weighting strategies: equal weighting as well as proportional to lexical similarity of the pivot to (i) source, (ii) target, (iii) average of (i) and (ii) (see Table 6 for results).

For each weighting strategy, **the interpolated system outperformed not just the individual pivot systems, but also the direct translation system.** We see more than 1.5 BLEU point improvement over the best single pivot model. Previous studies have shown that word and morph level multiple pivot systems were not able to outperform the direct system, possibly due to the effect of sparsity on triangulation (More et al., 2015; Dabre et al., 2015). Thus, it is remarkable that **multi-linguality could help overcome the lack of a direct translation system between the two languages.** Once the ill-effects of data sparsity are reduced due to the use of OS level pivot, multiple pivot languages can maximize translation performance because: (i) they bring in more translation options, and (ii) they improve the estimates of feature values with evidence from multiple languages.

We observe that equal weighting as well as assigning interpolation weights proportional to the average lexical similarity of the pivot to source and target are the best interpolation strategies.

Interpolation	mar-ben	mal-hin
equal	24.17	19.34
source	24.25	19.33
target	24.19	19.39
average	24.34	19.36
all_interpolate	24.41	19.44

Table 7: Results: Augmentation of direct system with pivot systems (%BLEU)

Pivot	BLEU	LCSR Similarity of pivot with		
		src (1)	tgt (2)	ave (1,2)
Marathi-Bengali translation				
pan	22.07	47.18	44.58	45.88
guj	22.54	<u>54.09</u>	48.99	51.53
hin	22.92	53.01	<u>52.30</u>	<u>52.66</u>
Malayalam-Hindi translation				
tel	17.26	<u>39.18</u>	42.94	41.06
mar	17.52	33.56	53.01	43.29
guj	17.44	32.92	<u>58.86</u>	<u>45.89</u>

Table 8: Choice of pivot language. Closest languages to source, target and average underlined.

7.2 Augmenting direct translation with pivot translation

We augmented the direct translation system with each of the interpolated systems discussed in the previous section using equal-weighted interpolation. In addition, we also tried combining all the pivot systems and the direct system using equal-weighted interpolation (all_interpolate). The results are shown in Table 7. We observe that the augmented system further improved the translation accuracy by about 1 BLEU point over the direct system. **Thus, the use of all related languages via pivoting and interpolation gave the best translation between a language pair.**

7.3 Choice of pivot language

Previous studies on the choice of pivot language have been impeded by the morphological diversity of the pivot languages and morphologically poorer pivots tend to perform better (More et al., 2015; Dabre et al., 2015; Paul et al., 2013). Thus the varied levels of sparsity induced by morphological properties dictated the choice of pivot, rather than the intrinsic properties of the pivot language. Sparsity is not a major concern with OS level models, hence we are able to study the effect of language properties on the choice of the pivot language.

We studied if the lexical similarity of the pivot language to the source and/or target language had

IL	IL-hin-eng			eng-hin-IL		
	M_{piv}	OS_{piv}	W_{dir}	M_{piv}	OS_{piv}	W_{dir}
ben	14.40	14.82	14.47	10.72	10.50	12.86
guj	16.66	17.24	17.23	13.67	13.59	16.20
mar	15.48	15.72	14.78	9.93	9.98	10.32

Table 9: Results for translation involving unrelated languages via related pivot (%BLEU). The translation is between English and an Indian language (IL) with a related Indian language (Hindi) as pivot. M_{piv} and O_{piv} means that the IL-Hindi translation is at the morph and OS levels respectively. W_{dir} is a direct English-IL translation model.

an impact on the translation quality. We studied *mal-hin* translation and *mar-ben* translation for multiple pivot languages. Table 8 shows the translation accuracy for different pivot languages along with lexical similarity of the pivot language to the (i) source (ii) target and (iii) average of (i) and (ii). The choice of pivot language makes a limited but observable difference (within 1 BLEU point).

For Marathi-Bengali translation, Hindi is the most beneficial pivot and it is most similar to the target as well as equidistant from both source and target. For Malayalam-Hindi translation, Marathi is the most useful pivot, with Gujarati close behind. Gujarati is most similar to the target as well as equidistant from both source and target. Gujarati is more similar to Marathi.

These observations suggest that a pivot language which is either closer to the target language or equidistant from both source and target is more useful than having a pivot which is closer to the source language. This provides further evidence to the observations by Paul et al. (2013) that target language features are more important for “coherent” language pairs.

8 Translation involving an unrelated language

So far, we have reported our investigations related to scenario [A]. But Scenario [B] (translation between unrelated languages via a pivot related to source or target) is also a common situation. For instance, no parallel corpus may exist for Marathi-English translation, but Hindi-English parallel corpus may be available (Hindi and Marathi are related languages). In this case, we adopt the transfer approach to pivot translation, but **the two legs of the pivot translation use different translation units. Marathi-Hindi trans-**

lation occurs at the subword level, while Hindi-English translation occurs at the word level. In this section, we present the results of our investigations into scenario [B]. We experimented with Indian language to English and vice versa via Hindi as pivot. We trained our Hindi-English translation models on the ILCI corpus. For English-Indian language (IL) translation, rule-based source reordering (Ramanathan et al., 2008) was used to overcome the structural divergence between English and Indian languages (English is an SVO languages and Indian languages are SOV).

Table 9 shows the results for the proposed approach for Indian Language-Hindi-English translation and vice versa. It also shows a comparison with word level direct translation results. For IL-hin-eng translation, the pivot system using OS model for IL-hin is better than the one using a morph model for IL-hin as well as a direct word level IL-eng translation model. However, the gains are minor compared to the gains we observed when all languages were related. For eng-hin-IL translation, using an OS level IL-hin pivot offers no advantage over a morph level IL-hin pivot. Possibly, the OS model is sensitive to word order errors made by the eng-hin leg of the pivot translation.

To summarize, the use of OS level model as one leg of a transfer-based pivot translation has only a minor benefit for translation between unrelated languages through a pivot related to either source or pivot. Better solutions need to be investigated.

9 Conclusion & Future Work

We investigated the use of pivot translation for translation involving related languages when direct parallel corpora are not available. We show that pivot translation between related languages can rival or outperform direct translation if **subword level translation** is done using **multiple related pivot languages**. Subword units make this possible by using lexical similarity and reducing losses in pivoting (owing to small vocabulary).

Our observations also hold lessons for the design of multilingual translation systems. It is better to invest in building/mining a parallel corpus through a related pivot (preferably closer to the target) than an unrelated pivot. When building translation systems between a *lingua franca* and an unrelated language, building/mining parallel corpora between the *lingua franca* and the related lan-

guages is not the best choice. Rather, it is better to consider one of the related languages as a pivot language. These design decisions can ensure better translation accuracy with lower investment in the development of parallel corpora.

Our results show that better methods are needed for pivot translation involving unrelated languages. The major problem current methods face is data sparsity resulting from word/morpheme level translation. Subword level translation between arbitrary languages is viable in the neural MT framework - hence, NMT may be a good research direction to tackle this translation scenario.

References

- Jose B Marino Adrià De Gispert. 2006. Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In *In Proc. of 5th International Conference on Language Resources and Evaluation (LREC)*.
- Pushpak Bhattacharyya, Mitesh Khapra, and Anoop Kunchukuttan. 2016. Statistical machine translation between related languages. www.cfilt.iitb.ac.in/publications/naacl-tutorials. NAACL Tutorials.
- Arianna Bisazza, Nick Ruiz, Marcello Federico, and FBK-Fondazione Bruno Kessler. 2011. Fill-up versus interpolation methods for phrase-based smt adaptation. In *IWSLT*. pages 136–143.
- Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, and Daniel Zeman. 2014. HindEnCorp – Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*.
- Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *ACL*.
- Raj Dabre, Fabien Cromieres, Sadao Kurohashi, and Pushpak Bhattacharyya. 2015. Leveraging small multilingual corpora for smt using many pivot languages. In *HLT-NAACL*. pages 1192–1202.
- Nadir Durrani, Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2010. Hindi-to-Urdu machine translation through transliteration. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.
- Murray B Emeneau. 1956. India as a linguistic area. *Language*.
- Girish Nath Jha. 2012. The TDIL program and the Indian Language Corpora Initiative. In *Language Resources and Evaluation Conference*.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*.
- Grzegorz Kondrak. 2005. Cognates and word alignment in bitexts. *Proceedings of the tenth machine translation summit (mt summit x)* pages 305–312.
- Anoop Kunchukuttan and Pushpak Bhattacharyya. 2016a. Faster decoding for subword level phrase-based smt between related languages. In *Third Workshop on NLP for Similar Languages, Varieties and Dialects*.
- Anoop Kunchukuttan and Pushpak Bhattacharyya. 2016b. Learning variable length units for SMT between related languages via byte pair encoding. *ArXiv e-prints*, arxiv:1610.06510.
- Anoop Kunchukuttan and Pushpak Bhattacharyya. 2016c. Orthographic syllable as basic unit for smt between related languages. In *Empirical Methods in Natural Language Processing*.
- Anoop Kunchukuttan, Ratish Pudupully, Rajen Chatterjee, Abhijit Mishra, and Pushpak Bhattacharyya. 2014. The IIT Bombay SMT System for ICON 2014 Tools contest. In *NLP Tools Contest at ICON 2014*.
- I Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Third Workshop on Very Large Corpora*.
- I Dan Melamed. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics* 25(1):107–130.
- Rohit More, Anoop Kunchukuttan, Raj Dabre, and Pushpak Bhattacharyya. 2015. Augmenting pivot based smt with word segmentation. In *International Conference on Natural Language Processing*.
- Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*.
- Michael Paul, Andrew Finch, and Eiichiro Sumita. 2013. How to choose the best pivot language for automatic translation of low-resource languages. *ACM Transactions on Asian Language Information Processing (TALIP)* 12(4):14.

- Uwe Quasthoff, Matthias Richter, and Christian Bie-
mann. 2006. Corpus portal for search in monolin-
gual corpora. In *Proceedings of the fifth interna-
tional conference on language resources and evalu-
ation*.
- Ananthakrishnan Ramanathan, Jayprasad Hegde,
Ritesh M Shah, Pushpak Bhattacharyya, and
M Sasikumar. 2008. Simple syntactic and morpho-
logical processing can help english-hindi statistical
machine translation. In *IJCNLP*. pages 513–520.
- Loganathan Ramasamy, Ondřej Bojar, and Zdeněk
Žabokrtský. 2012. Morphological Processing for
English-Tamil Statistical Machine Translation. In
*Proceedings of the Workshop on Machine Transla-
tion and Parsing in Indian Languages*.
- Calvin R Rensch. 1992. Calculating lexical similarity.
In Eugene H. Casad, editor, *Windows on bilingual-
ism*, Summer Institute of Linguistics and the Univer-
sity of Texas at Arlington.
- TAUS. 2013. Google Translate.
https://www.taus.net/knowledgebase/index.php?title=Google_Translate.
Accessed: 2017-02-04.
- Jörg Tiedemann. 2009. Character-based PBSMT for
closely related languages. In *Proceedings of the
13th Conference of the European Association for
Machine Translation*.
- Jörg Tiedemann. 2012. Character-based pivot transla-
tion for under-resourced languages and domains. In
EACL.
- Jörg Tiedemann and Preslav Nakov. 2013. Analyzing
the use of character-level translation with sparse and
noisy datasets. In *RANLP*.
- Masao Utiyama and Hitoshi Isahara. 2007. A compari-
son of pivot methods for phrase-based statistical ma-
chine translation. In *HLT-NAACL*. pages 484–491.
- David Vilar, Jan-T Peter, and Hermann Ney. 2007. Can
we translate letters? In *Proceedings of the Second
Workshop on Statistical Machine Translation*.
- Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko
Kurimo, et al. 2013. Morfessor 2.0: Python im-
plementation and extensions for morfessor baseline.
Technical report, Aalto University.
- Hua Wu and Haifeng Wang. 2007. Pivot language ap-
proach for phrase-based statistical machine transla-
tion. *Machine Translation* 21(3):165–181.